Разведанализ данных

В нашем исследовании мы используем данные по **50 штатам США с 2010 по 2019 год.**

Чтобы изучить экономические условия штата, мы включили такие характеристики как количество совершенных насильственных преступлений, уровень бедности и безработицы, а также доход и реальный ВВП.

Данные были собраны с разных ресурсов, так что требуют некоторых преобразований для дальнейшего анализа.

	Year	State	Vcrime	PR	UR	PIPC	RGDP
	<int>	<chr>	<dbl>	<dbl>	<dbl>	<int>	<dbl>
1	2010	Alabama	449.8	0.191	0.10387181	33963	184702.4
2	2010	Alaska	633.0	0.111	0.08162109	49654	54601.5
3	2010	Arizona	408.3	0.175	0.10299779	33848	260307.1
4	2010	Arkansas	517.7	0.188	0.07873529	32367	105662.0
5	2010	California	472.0	0.158	0.12476958	43249	2036015.0
6	2010	Colorado	337.8	0.131	0.09158730	40785	267858.9

6 rows

Рассмотрим обощающие данные по всем переменным:

Переменная Year

Показывает год, за который были совершены преступления, является числовой.

Мы взяли данные за 2010-2019 года, за 2020 не брали в целом, так как некоторые из необходимых значений будут известны лишь к концу 2021 - началу 2022 года из-за особенностей сбора статистики в отдельных штатах.

Переменная State

```
   Length     Class      Mode 
      500 character character 
```

Текстовая переменная, обозначающая название штата.

```
 [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
 [5] "California"     "Colorado"       "Connecticut"    "Delaware"
 [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
[13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
[17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
[21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
[25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
[29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
[33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
[37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
[41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
[45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
[49] "Wisconsin"      "Wyoming"
```

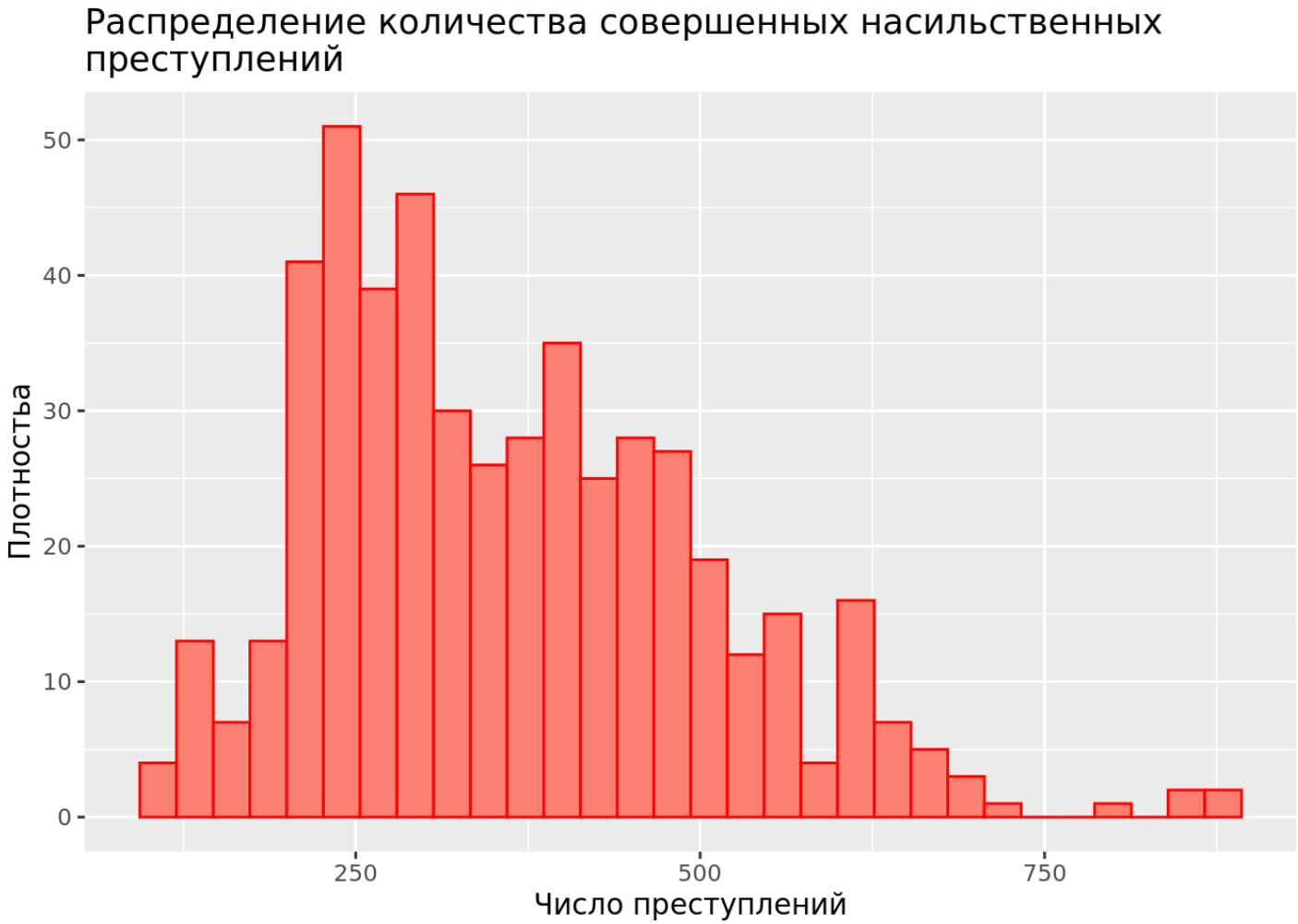
Имеет 50 уникальных значений.

Переменная Vcrime

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  112.0   252.4   339.0   362.9   451.8   885.0
```

Числовая переменная, указывающий количество совершенных насильственных преступлений (violent crimes) на 100 тысяч населения.

Значения колеблются в диапазоне от 112 до 885, где среднее (362) и медианное (339) близки друг к другу.



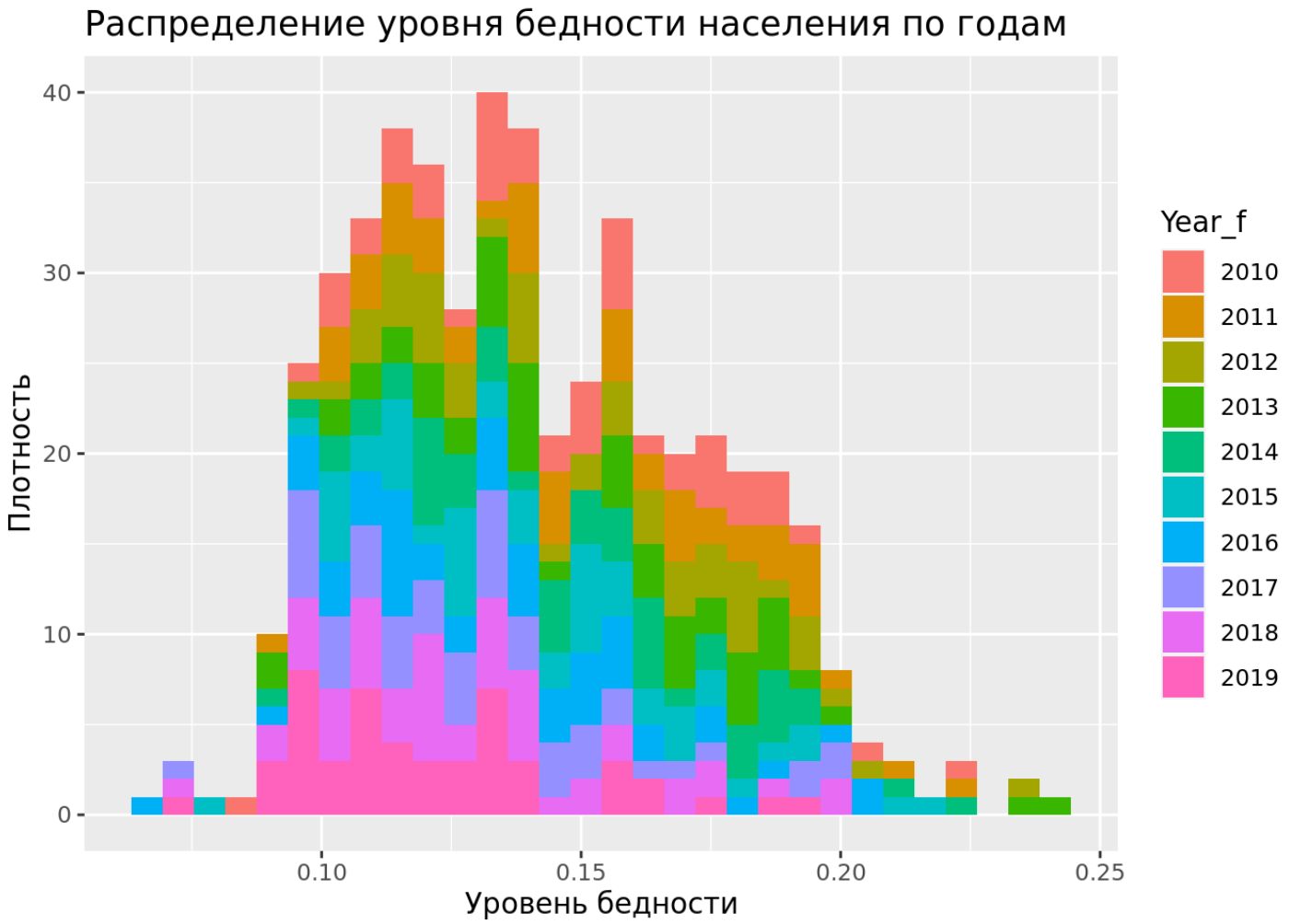
На графике видно, что количество регистрируемых наблюдений в целом находится на достаточно высоком уровне, что еще раз подтверждает выводы СМИ о существующей проблеме с насильственными преступлениями на территории Соединенных Штатов Америки. Статистический службы также зафиксировали и несколько экстремально высоких значений за все время наблюдения (почти 1000 преступлений). Подобные выбросы можно объяснить снижением затрат на финансирование правоохранительных органов, усилением расслоения общества, экономическими шоками. Следует заметить, что это могло произойти и вследствие ошибок при фиксации и регистрации насильственных преступлений.

Переменная PR

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0690  0.1150  0.1360  0.1405  0.1640  0.2440 
```

Poverty rate - числовая переменная, означающая уровень бедности в долях, то есть от 0 до 1.

В нашем случае минимальное значение 0.069, а максимальное 0.244

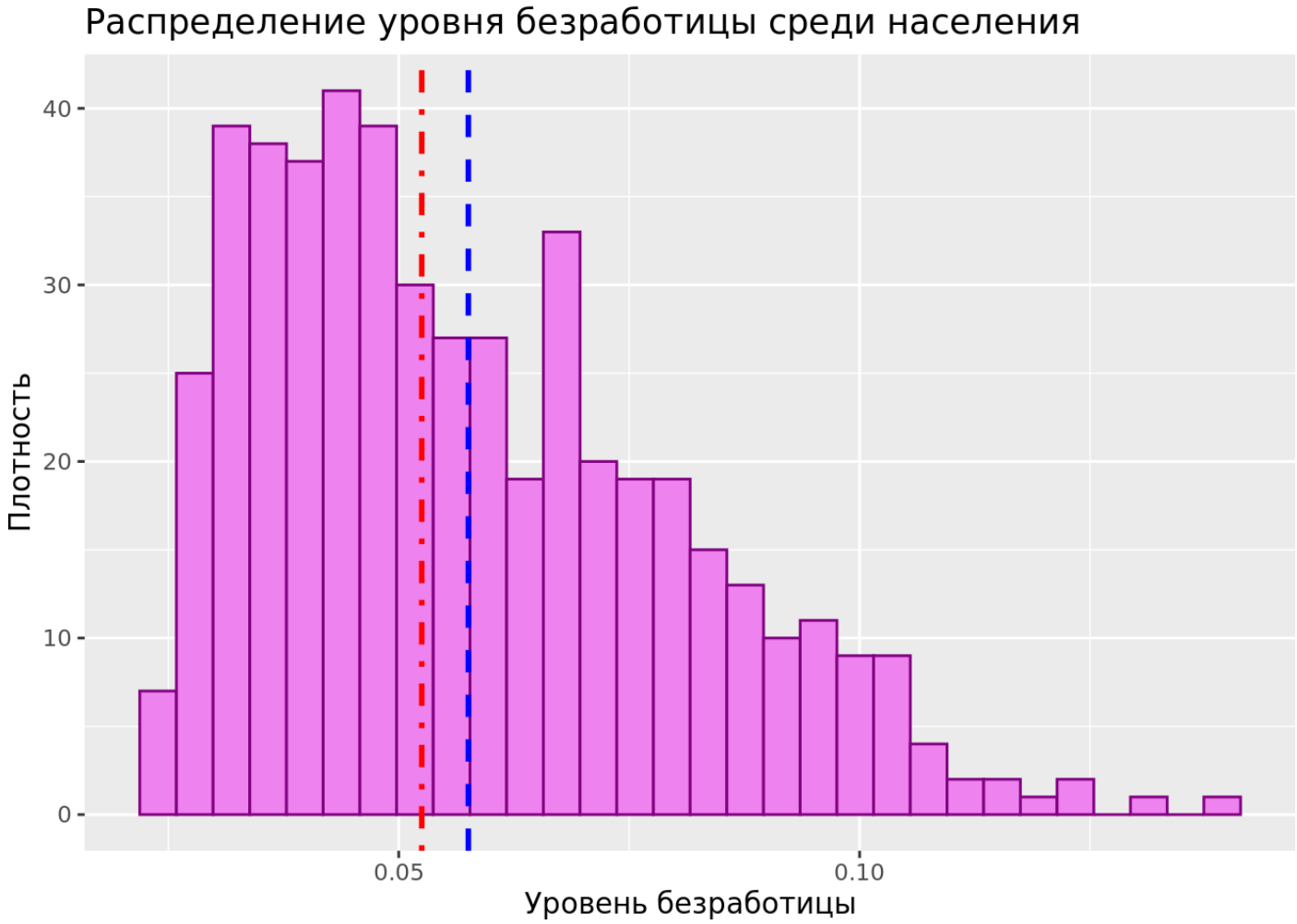


Средний зафиксированный за весь период наблюдения уровень бедности по всем штатам показывает неплохие - относительно мировых - значения. По графику можно заметить, что уровень бедности падает с каждым годом, что несомненно служит хорошим показателем экономической эффективности штата. Тем не менее, стоит упомянуть и крайне высокие значения, близкие к 0,25. Эти показатели были зафиксированы в начале десятых годов (2012 и 2013). Сложно однозначно сказать, что послужило причиной подобных выбросов.

Переменная UR

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.02215 0.03999 0.05250 0.05755 0.07225 0.13760 
```

Unemployment rate - уровень безработицы в штатах в долях. Значения колеблются от 2,2% до 13,7%



На графике видно, что среднее значение и медиана колеблются около 5% уровня безработицы, как было в таблице. Кроме того, можно заметить, что показателей выше 11% довольно мало, по сравнению с остальными.

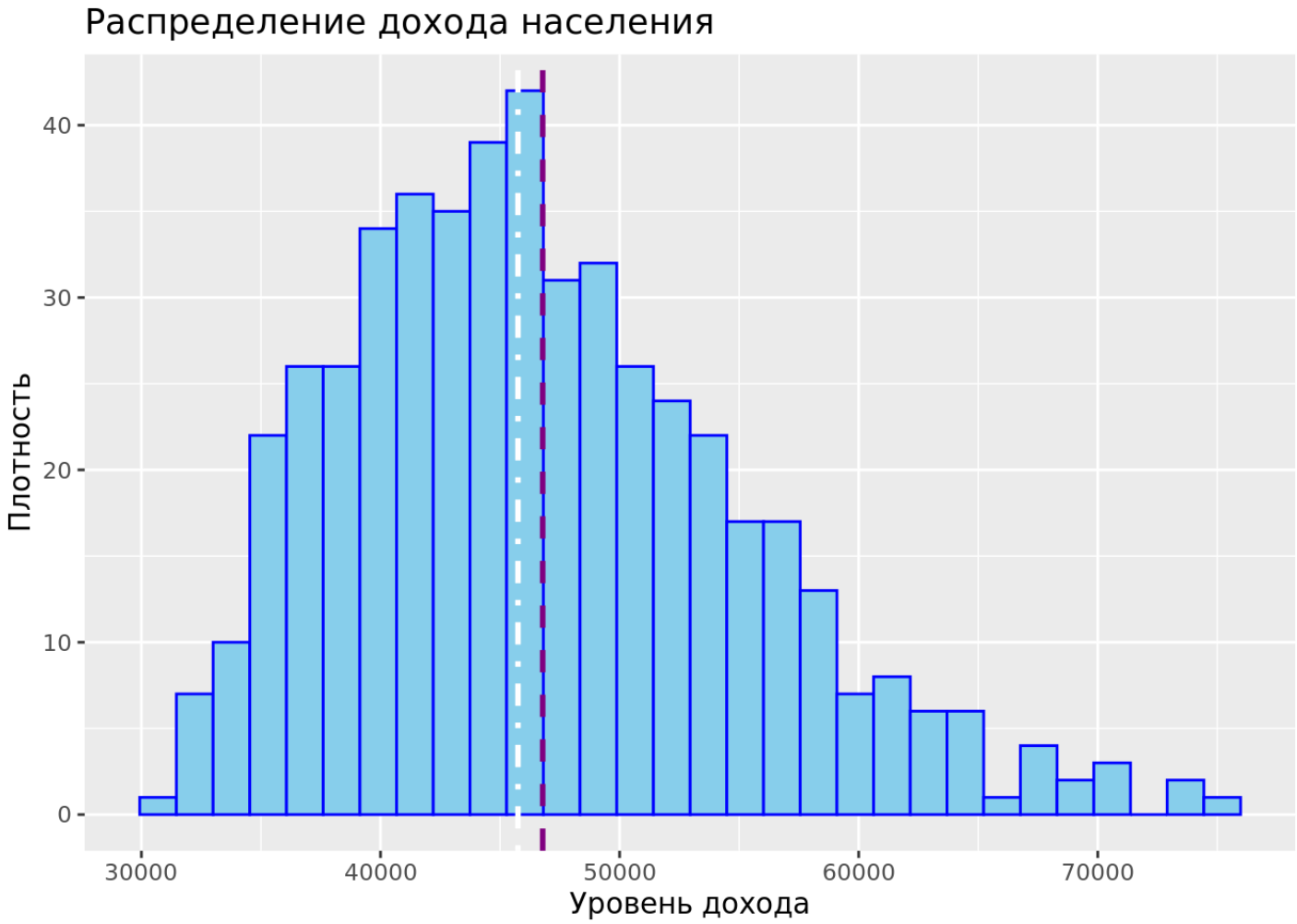
Также видно, что в среднем в Соединенных Штатах Америки сохраняется достаточно низкий уровень безработицы, что может указывать на эффективность работы служб занятости и наличии свободных рабочих мест. Выбросы на графике можно объяснить локальными коллапсами или, например, длительным восстановлением после кризиса 2007-2009 годов

Переменная PIPC

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  31284   40656   45751   46785   52224   75794 
```

Personal income per capita - числовое значение, выражающее доход на душу населения в долларах.

Максимальное значение (75794) более чем в два раза больше минимального (31284), хотя среднее и медианное значения чуть больше 45000. Получается, что значений, близких к максимальному, довольно мало.

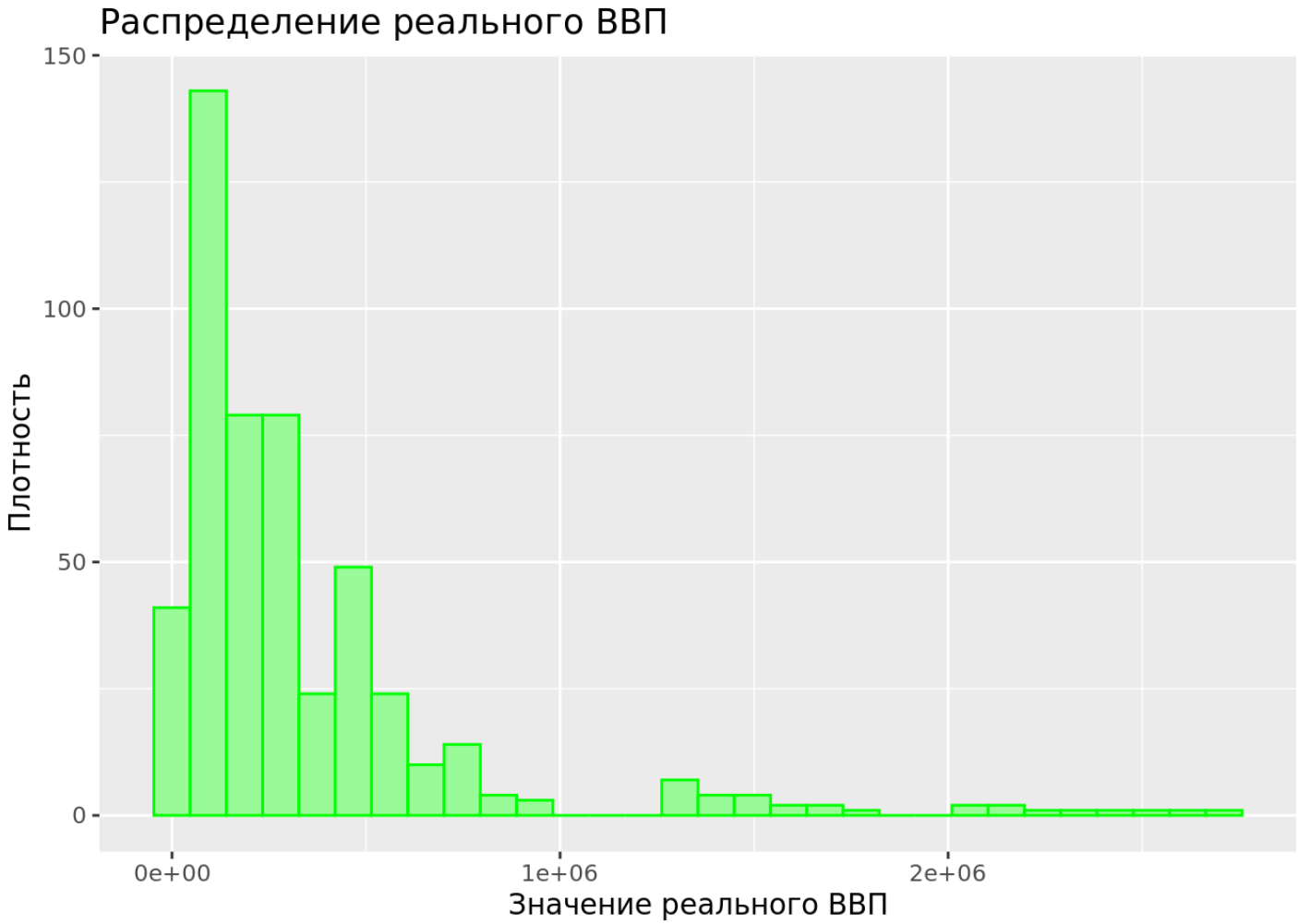


Выбросы на графике можно объяснить спецификой данного показателя: он фиксирует все доходы человека (зарплата, доходы от аренды, дивиденды, трансферты и другие виды). Это приводит к следующей ситуации: в маленьких штатах показатель PIPC может быть либо слишком высоким, либо слишком низким из-за небывалой урожайности (особенно актуально для аграрно-зависимых штатов), природной катастрофы или крупного экономического проекта. В то же время дополнительное население (например, студенты колледжей из других штатов) может занижать показатели PIPC.

Переменная RGDP

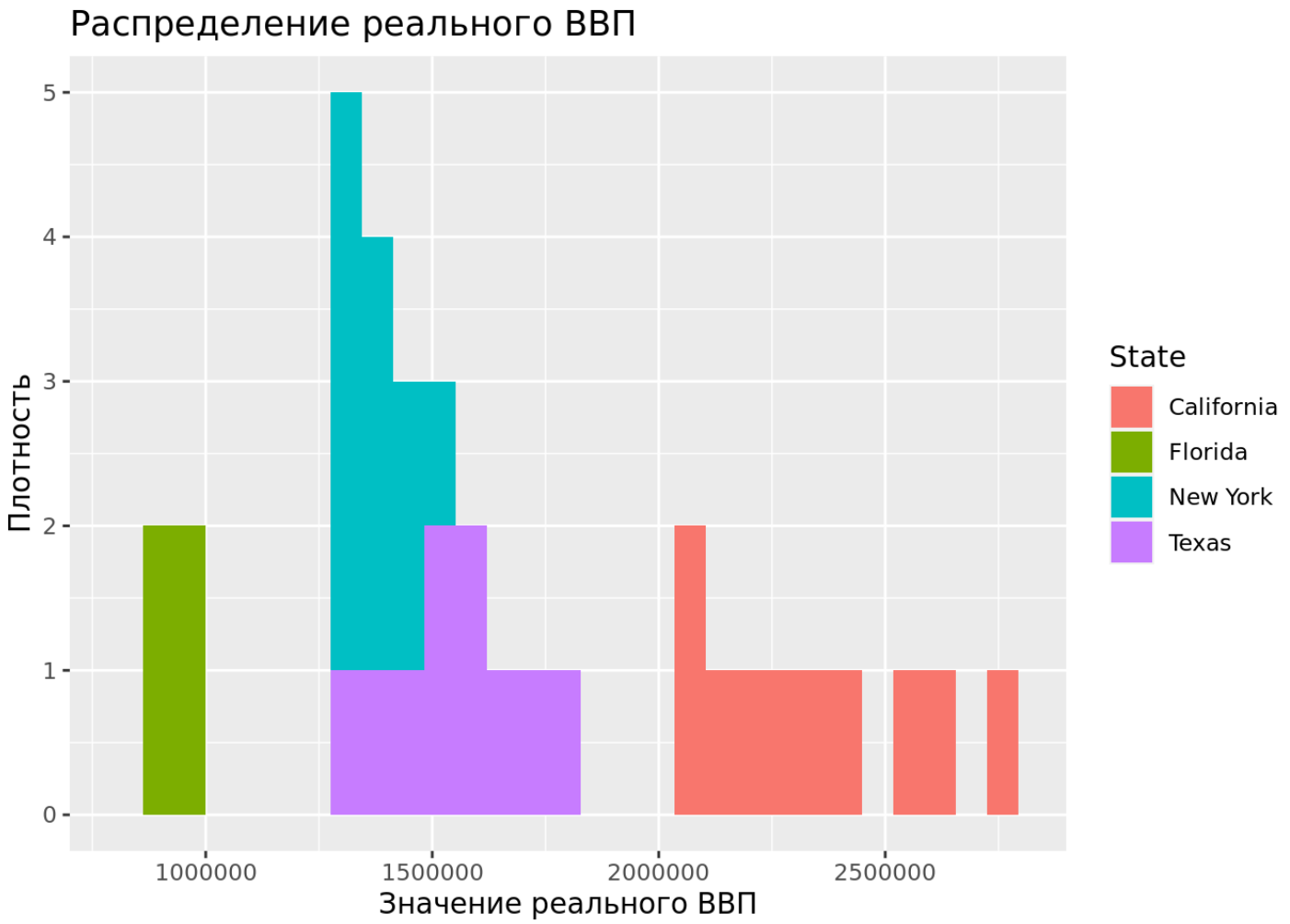
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  28404   76812  198042  339150  441338 2739343 
```

Real GDP - числовое значение реального ВВП штата. За 10 лет он колеблется от 28404 до 2739343.



По графику видно, что некоторые значения сильно выбиваются по значению ВВП в большую сторону.

Чтобы понять, как именно выглядит выброс, рассмотрим часть графика ближе:



Таким образом мы видим, что выбивается в большей степени штат Калифорнии.

Посмотрим данные дополнительно:

	Year	State	Vcrime	PR	UR	PIPC	RGDP	Year_f
	<int>	<chr>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<fctr>
1	2019	California	447.0	0.118	0.04150195	64513	2739343	2019
2	2018	California	447.0	0.128	0.04264428	61663	2643576	2018
3	2017	California	445.0	0.132	0.04834575	58942	2541769	2017
4	2016	California	426.0	0.144	0.05505001	56667	2427895	2016
5	2015	California	402.0	0.153	0.06251195	54632	2357453	2015
6	2014	California	423.1	0.164	0.07561169	51332	2256055	2014
7	2013	California	423.1	0.167	0.09035860	48549	2179229	2013
8	2012	California	411.1	0.169	0.10515859	48154	2113096	2012
9	2011	California	440.6	0.167	0.11876914	45574	2063828	2011
10	2010	California	472.0	0.158	0.12476958	43249	2036015	2010

1-10 of 15 rows Previous **1** 2 Next

Таким образом мы видим, что штат Калифорния отличается от других штатов тем, что реальный ВВП с 2010 по 2019 год было выше 2000000, а у остальных штатов значительно ниже. Нельзя назвать это выбросом данных, так как все значения принадлежат одному штату за все года, то есть это реально возможные значения.

Нетрудно объяснить подобное отклонение реального ВВП: Калифорния - крупнейшая экономика в рамках Соединенных Штатов Америки, пятая по показателям ВВП даже среди стран. Одними из драйверов подобного успеха являются численность населения и диверсификация экономики (финансовый сектор и недвижимость, информационная отрасль, производственный сектор вносят наибольший вклад).

Общий вывод

Нам удалось собрать достаточно большой датасет с данными за обширный период времени, на основе которого мы сможем проверить нашу гипотезу о наличии зависимости между экономическими условиями штата и количеством насильственных преступлений в нем. В рамках разведанализа нам удалось провести чистку данных и убедиться в правильности выбора переменных в рамках нашего исследования.